Role of Scene text in Image Semantics

A Thesis submitted by
Arka Ujjal Dey

in partial fulfillment of the requirements for the award of the degree of
Doctor of Philosophy

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology Jodhpur
Department of Computer Science and Engineering
August 2022

Declaration

I hereby declare that the work presented in this Thesis titled *Role of Scenetext in Image semantics* submitted to the Indian Institute of Technology Jodhpur in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out under the supervision of Dr. Gaurav Harit. The contents of this thesis in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Arka Ujjal Dey
P15CS001

Certificate

This is to certify that the thesis titled *Role of Scenetext in Image Semantics*, submitted by *Arka Ujjal Dey (P15CS001)* to the Indian Institute of Technology Jodhpur for the award of the degree of *Doctor of Philosophy*, is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Gaurav Harit
Ph.D. Thesis Supervisor

Acknowledgments

I thank Asimov and ELIZA for making me think how similar we are to automatons. And to my family and their appetite for academics, making me believe research is 'cool'. This work would not have been possible without Dr.Harit, those sustained support has helped it take shape. A special mention to my colleagues, in IIT Jodhpur and CVC Barcelona, for the healthy debates.

List of Figures

List of Tables

List of Symbols

A	Attention Matrix
$\mathbf{r}_{i,k}$	gives the similarity between a scene text word $t_i^{'}$ and anchor A_k
$\mathbf{t}_i^{'}$	scene text feature for i^{th} word
$\mathbf{t}_k$	scene text features in terms of k^{th} anchor
$\mathbf{v}_i^{res}$	ResNet-152 Convolutional Features for the i^{th} object
$\mathbf{v}_i$	Visual Features for the i^{th} object
W	learned projection matrix
α_{ij}	the attention weight representing the importance of node j to node i
$\alpha_{\mathbf{z}}^{prv}$	The weighing factor for the rest of the nodes in the final representation of the decoder inpu
d_{cs}	Refers to the cosine distance
d_L	Refers to the edit distance
D	Number of Decoding Steps
H	Number of Vocabulary objects in classifier
L	Number of Question Words
M	Number of Visual Objects
N	Number of Scene text Objects
$\mathbf{r}_j$	validity score for j^{th} candidate knowledge $\mathbf{S}_{i_j} \in \mathbf{S}_i$
$\mathbf{S}_i$	Set of candidate knowledge facts for scene text word $\mathbf{t}_i$
$\mathbf{x}_p^{\mathbf{c}}$	Feature representation of p^{th} context object
$\mathbf{x}_m^{\mathbf{obj}}$	Feature representation of m^{th} visual object
$\mathbf{x}_n^{\mathbf{ocr}}$	Feature representation of n^{th} scene text object
$\mathbf{x}_n^{\mathbf{knw}}$	Feature representation of the knowledge fact corresponding to n^{th} scene text object
$\mathbf{x}_d^{\mathbf{prv}}$	Feature representation at the d^{th} step of the previous-step object
$\mathbf{x}_l^{\mathbf{ques}}$	Feature representation of l^{th} question word
$\mathbf{y}_d^{\mathbf{ocr}}$	Pointer network score for OCR objectes at the d^{th} step
$\mathbf{y}_d^{\mathbf{voc}}$	classifier score for answer vocabulary objects at the d^{th} step
$\mathbf{z}_m^{\mathbf{obj}}$	Transformer output of m^{th} visual object
$\mathbf{z}_n^{\mathbf{ocr}}$	Transformer output of n^{th} scene text object
$\mathbf{z}_n^{\mathbf{knw}}$	Transformer output of the knowledge fact corresponding to n^{th} scene text object
$\mathbf{z}_d^{\mathbf{prv}}$	Transformer output of d^{th} previous decoding step
$\mathbf{z}_l^{\mathbf{ques}}$	Transformer output of l^{th} question word

List of Abbreviations

ANLS	Averaged Normalized Levenshtein Similarity
BERT	Bidirectional Encoder Representations from Transformers
BiLM	Bi-directional Language Model
BOW	Bag Of Words
CE	Cross Entropy
CNN	Convolutional Neural Network
CVC	Computer Vision Center
EKB	External Knowledge Base
EKTVQA	External Knowledge-enabled Text Visual Question Answering
FC	Fully Connected
FCNN	Fully Connected Neural Network
F-RCNN	Faster R-CNN
GAN	Generative Adversarial Network
GAT	Graph Attention Layers
GKB	Google Knowledge Base
GCN	Graph Convolution Network
GPU	Graphics Processing Unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ITM	Image Text (contrastive) Matching
LSTM	Long Short Term Memory
M4C	Multi-modal Multi-Copy Mesh
MLM	Masked Language Modeling
MMHS150K	Multimedia Hate Speech Dataset
MSE	Mean Squared Error
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLS	Normalized Levenshtein Similarity
OCR	Optical Character Recognition
PHOC	Pyramid Histogram of Characters
R@K	Recall at K
RPP	Relative (spatial) Position Prediction
ReLU	Rectified Linear Unit
SA-M4C	Spatially Aware M4C
SGD	Stochastic Gradient Descent
SMA	Structured Multi-modal Attention
SVM	Supper Vector Machine
TAP	Text aware Pretraining
Text-VQA	Text Visual Question Answering
Text-KVQA	Text Knowledge Visual Question Answering
TF-IDF	Term Frequency-Inverse Document Frequency
TVQA	Our baseline model for Text-VQA
Vocab UB	Vocabulary Upper bound referred
VLN	Vision and Language Navigation

VQA Visual Question Answering
VGG Visual Geometry Group